1 Parameter estimation

Statisticians do it when it counts.

1.1 What is Statistics?

***Statistics** is a collection of procedures and principles for gaining and processing information in order to make decisions when faced with uncertainty.*

This course is concerned with "Mathematical Statistics", i.e., mathematical ideas and tools that are used by statisticians to analyse data. We will study techniques for estimating parameters, fitting models, and testing hypotheses. However, as we study these techniques we should not lose sight of the fact that a practicing statistician needs more than simply a knowledge of mathematical techniques. The collection and interpretation of data is a subtle art. It requires common sense. It can sometimes raise philosophical questions. Although this course is primarily concerned with mathematical techniques, I will try — by means of examples and digressions — also to introduce you to some of the non-mathematical aspects of Statistics.

Statistics is concerned with *data analysis*: using data to make inferences. It is concerned with questions like 'what is this data telling me?' and 'what does this data suggest it reasonable to believe?' Two of its principal concerns are **parameter estimation** and **hypothesis testing**.

Example 1.1 *Suppose we wish to estimate the proportion p of students in Cambridge who have not showered or bathed for over a day.*

This is poses a number of questions. Who do we mean by students? Suppose time is limited and we can only interview 20 students in the street. Is it important that our survey be 'random'? How can we ensure this? Will students we question be embarrassed to admit if they have not bathed? And even if we can get truthful answers, will we be happy with our estimate if that random sample turns out to include no women, or if it includes only computer scientists?

Suppose we find that 5 have not bathed for over a day. We might estimate p by $\hat{p} = 5/20 = 0.25$. But how large an error might we expect $\hat{p}$ to have?

Many families of probability distributions depend on a small number of parameters; for example, the Poisson family depends on a single parameter λ and the Normal family on two parameters μ and σ. Unless the values of the parameters are known in advance, they must be estimated from data. One major themes of mathematical statistics is the theory of **parameter estimation** and its use in fitting probability distributions to data. A second major theme of Statistics is **hypothesis testing**.

Example 1.2 *A famous study investigated the effects upon heart attacks of taking an aspirin every other day.* The results after 5 years were

Condition	**Heart attack**	**No heart attack**	**Attacks per 1000**
Aspirin	104	10,933	9.42
Placebo	189	10,845	17.13

What can make of this data? Is it evidence for the hypothesis that aspirin prevents heart attacks?

The aspirin study is an example of a *controlled experiment*. The subjects were doctors aged 40 to 84 and none knew whether they were taking the aspirin or the placebo. Statistics is also concerned with analysing data from *observational studies*. For example, most of us make an intuitive statistical analysis when we use our previous experience to help us choose the shortest checkout line at a supermarket.

The data analysis of observational studies and experiments is a central component of **decision-making**, in science, medicine, business and government.

By the way: **data** is a plural noun referring to a collection of numbers or other pieces of information to which meaning has been attached.

The numbers 1.1, 3.0, 6.5 are not necessarily data. They become so when we are told that they are the muscle weight gains in kg of three athletes who have been trying a new diet.

1.2 RVs with values in $\mathbb{R}^n$ or $\mathbb{Z}^n$

In Statistics, our data are modelled by a vector of random variables

$$X = (X_1, X_2, \dots, X_n)$$

where X_i takes values in $\mathbb{Z}$ or $\mathbb{R}$.

To succeed in this course you should brush up on your knowledge of basic probability: of key distributions and how to make calculations with random variables. Let us review a few facts.

When our sample space Ω (a set of possible outcomes) is discrete (finite or countably infinite) we have a random variable (**RV**) X with values in $\mathbb{Z}$:

$$X : \Omega \to \mathbb{Z}.$$

RVs can also take values in $\mathbb{R}$ rather than in $\mathbb{Z}$ and the sample space Ω can be uncountable.

$$X : \Omega \to \mathbb{R}.$$

Since the outcome ω, $\omega \in \Omega$, is random, X is a function whose value, $X(\omega)$, is also random. E.g., to model the experiment of tossing a coin twice we might take $\Omega = \{hh, ht, th, th\}$. Then X might be the total number of heads.

In both cases the **distribution function** F_X of X is defined as:

$$F_X(x) := \mathbb{P}(X \le x) = \sum_{\{\omega:\, X(\omega)\le x\}} \mathbb{P}(\omega).$$

In the discrete case the **probability mass function** (pmf) f_X of X is

$$f_X(k) := \mathbb{P}(X = k), \qquad k \in \mathbb{Z}.$$

So

$$\mathbb{P}(X \in A) = \sum_{x\in A} f_X(x), \quad A \subseteq \mathbb{Z}.$$

In the continuous case we have the **probability density function** (pdf) f_X of X. In all cases we shall meet, X will have a piecewise smooth pdf such that

$$\mathbb{P}(X \in A) = \int_{x\in A} f_X(x)\,dx, \quad \text{for } \textit{nice} \text{ (measurable) subsets } A \subseteq \mathbb{R}.$$

Expectation of X: In the discrete case

$$\mathbb{E}(X) := \sum_{\omega\in\Omega} X(\omega)\mathbb{P}(\omega) = \sum_{k\in\mathbb{Z}} k\,\mathbb{P}(X = k),$$

the first formula being the real definition. In the continuous case the calculation

$$E(X) = \int_\Omega X(\omega)\,\mathbb{P}(d\omega)$$

needs measure theory. However,

$$f_X(x) = \frac{d}{dx}F_X(x) \quad \text{except perhaps for finitely many } x.$$

Measure theory shows that for any nice function h on $\mathbb{R}$,

$$\mathbb{E}\,h(X) = \int_\mathbb{R} h(x)f_X(x)\,dx\,.$$

Variance of X: If $\mathbb{E}(X) = \mu$, then

$$\mathrm{var}(X) = \mathbb{E}(X-\mu)^2 = \mathbb{E}(X^2) - \mu^2.$$

1.3 Some important random variables

(a) We say that X has the **binomial distribution** $B(n,p)$, and write $X \sim B(n,p)$, if

$$\mathbb{P}(X = k) = \begin{cases} \binom{n}{k} p^k(1-p)^{n-k} & \text{if } k \in \{0,\dots,n\},\\ 0 & \text{otherwise.}\end{cases}$$

Then $\mathbb{E}(X) = np$, $\mathrm{var}(X) = np(1-p)$. This is the distribution of the number of successes in n independent trials, each of which has probability of success p.

(b) We say that X has the **Poisson distribution** with parameter λ, and write $X \sim P(\lambda)$, if

$$\mathbb{P}(X = k) = \begin{cases} e^{-\lambda}\lambda^k/k! & \text{if } k \in \{0,1,2,\dots\},\\ 0 & \text{otherwise.}\end{cases}$$

Then $\mathbb{E}(X) = \mathrm{var}(X) = \lambda$. The Poisson is the limiting distribution of $B(n,p)$ as $n \to \infty$ and $p \to 0$ with $\lambda = np$.

(c) We say that X is **standard normal**, and write $X \sim N(0,1)$, if

$$f_X(x) = \varphi(x) := \frac{1}{\sqrt{2\pi}}\exp(-x^2/2), \qquad -\infty \le x \le \infty.$$

Then

$$F_X(x) = \int_{-\infty}^x f_X(y)\,dy = \Phi(x) := \int_{-\infty}^x \varphi(y)\,dy.$$

Then $\mathbb{E}(X) = 0$, $\mathrm{var}(X) = 1$. Φ and φ are standard notations.

(d) We say that X is **Normal** with mean μ and variance σ^2 and write $X \sim N(\mu,\sigma^2)$ if

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad -\infty \le x \le \infty.$$

Then $\mathbb{E}(X) = \mu$, $\mathrm{var}(X) = \sigma^2$.

(e) We say that X is **uniform** on $[a,b]$, and write $X \sim U[a,b]$, if

$$f_X(x) = \frac{1}{b-a}, \qquad x \in [a,b].$$

Then $\mathbb{E}(X) = \frac{1}{2}(a+b)$, $\mathrm{var}(X) = \frac{1}{12}(b-a)^2$.

1.4 Independent and IID RVs

Random variables $X_1, \dots, X_n$ are called **independent** if for all $x_1, \dots, x_n$

$$\mathbb{P}(X_1 \le x_1; \dots; X_n \le x_n) = \mathbb{P}(X_1 \le x_1) \cdots \mathbb{P}(X_n \le x_n).$$

IID stands for independent identically distributed. Thus if $X_1, X_2, \dots, X_n$ are IID RVs, then they all have the same distribution function and hence the same mean and same variance.

We work with the probability mass function (pmf) of X in $\mathbb{Z}^n$ or probability density function (pdf) of X in $\mathbb{R}^n$: In most cases, $X_1, \dots, X_n$ are independent, so that if $x = (x_1, \dots, x_n) \in \mathbb{R}^n$, then

$$f_X(x) = f_{X_1}(x_1) \cdots f_{X_n}(x_n).$$

1.5 Indicating dependence on parameters

If $X \sim N(\mu, \sigma^2)$, then we indicate the dependence of the pdf of X on μ and σ^2 by writing it as

$$f(x \mid \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Or if $X = (X_1, \dots, X_n)$, where $X_1, \dots, X_n$ are IID $N(\mu, \sigma^2)$, then we would have

$$f(x \mid \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{\|x-\mu\mathbf{1}\|^2}{2\sigma^2}\right)$$

where $\mu\mathbf{1}$ denotes the vector $(\mu, \mu, \dots, \mu)^\top$.

In general, we write $f(x \mid \theta)$ to indicate that the pdf depends on a parameter θ. θ may be a vector of parameters. In the above $\theta = (\mu, \sigma^2)^\top$. An alternative notation we will sometimes employ is $f_\theta(x)$.

The set of distributions with densities $f_\theta(x)$, $\theta \in \Theta$, is called a **parametric family**. E.g.,, there is a parametric family of normal distributions, parameterised by values of μ, σ^2. Similarly, there is a parametric family of Poisson distributions, parameterised by values of λ.

1.6 The notion of a statistic

A **statistic**, $T(x)$, is any function of the data. E.g., given the data $x = (x_1, \dots, x_n)$, four possible statistics are

$$\frac{1}{n}(x_1 + \cdots + x_n), \quad \max_i x_i, \quad \frac{x_1 + x_3}{x_n} \log x_4, \quad 1997 + 10 \min_i x_i.$$

Clearly, some statistics are more natural and useful than others. The first of these would be useful for estimating μ if the data are samples from a $N(\mu, 1)$ distribution. The second would be useful for estimating θ if the data are samples from $U[0, \theta]$.

1.7 Unbiased estimators

An **estimator** of a parameter θ is a function $T = T(X)$ which we use to estimate θ from an observation of X. T is said to be **unbiased** if

$$\mathbb{E}(T) = \theta.$$

The expectation above is taken over X. Once the actual data x is observed, $t = T(x)$ is the **estimate** of θ obtained via the estimator T.

Example 1.3 *Suppose $X_1, \dots, X_n$ are IID $B(1, p)$ and p is unknown. Consider the estimator for p given by $\hat{p}(X) = \bar{X} = \sum_i X_i/n$.* Then $\hat{p}$ is unbiased, since

$$\mathbb{E}\hat{p}(X) = \mathbb{E}\left[\frac{1}{n}(X_1 + \cdots + X_n)\right] = \frac{1}{n}(\mathbb{E}X_1 + \cdots + \mathbb{E}X_n) = \frac{1}{n}np = p\,.$$

Another possible unbiased estimator for p is $\tilde{p} = \frac{1}{3}(X_1 + 2X_2)$ (i.e., we ignore most of the data.) It is also unbiased since

$$\mathbb{E}\tilde{p}(X) = \mathbb{E}\left[\frac{1}{3}(X_1 + 2X_2)\right] = \frac{1}{3}(\mathbb{E}X_1 + 2\mathbb{E}X_2) = \frac{1}{3}(p + 2p) = p\,.$$

Intuitively, the first estimator seems preferable.

1.8 Sums of independent RVs

In the above calculations we have used the fact the expectation of a sum of random variables is the sum of their expectations. It is always true (even when $X_1, \dots, X_n$ are not independent) that

$$\mathbb{E}(X_1 + \cdots + X_n) = \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n),$$

and for linear combinations of RVs

$$\mathbb{E}(a_1 X_1 + \cdots + a_n X_n) = a_1 \mathbb{E}(X_1) + \cdots + a_n \mathbb{E}(X_n).$$

If $X_1, X_2, \dots, X_n$ are independent, then

$$\mathbb{E}(X_1 X_2 \dots X_n) = \mathbb{E}(X_1)\mathbb{E}(X_2) \cdots \mathbb{E}(X_n),$$
$$\operatorname{var}(X_1 + \cdots + X_n) = \operatorname{var}(X_1) + \cdots + \operatorname{var}(X_n),$$

and for linear combinations of independent RVs

$$\operatorname{var}(a_1 X_1 + \cdots + a_n X_n) = a_1^2 \operatorname{var}(X_1) + \cdots + a_n^2 \operatorname{var}(X_n).$$

1.9 More important random variables

(a) We say that X is **geometric** with parameter p, if

$$\mathbb{P}(X=k)=\begin{cases} p(1-p)^{k-1} & \text{if } k\in\{1,2,\dots\},\\ 0 & \text{otherwise.}\end{cases}$$

Then $E(X)=1/p$ and $\mathrm{var}(X)=(1-p)/p^2$. X is the number of the toss on which we first observe a head if we toss a coin which shows heads with probability p.

(b) We say that X is **exponential** with **rate** λ, and write $X\sim\mathcal{E}(\lambda)$, if

$$f_X(x)=\begin{cases} \lambda e^{-\lambda x} & \text{if } x>0,\\ 0 & \text{otherwise.}\end{cases}$$

Then $\mathbb{E}(X)=\lambda^{-1}$, $\mathrm{var}(X)=\lambda^{-2}$.

The geometric and exponential distributions are discrete and continuous analogues. They are the unique 'memoryless' distributions, in the sense that $\mathbb{P}(X\geq t+s\mid X\geq t)=\mathbb{P}(X\geq s)$. The exponential is the distribution of the time between successive events of a Poisson process.

(c) We say that X is **gamma**(n,λ) if

$$f_X(x)=\begin{cases} \lambda^n x^{n-1}e^{-\lambda x}/(n-1)! & \text{if } x>0,\\ 0 & \text{otherwise.}\end{cases}$$

X has the distribution of the sum of n IID RVs that have distribution $\mathcal{E}(\lambda)$. So $\mathcal{E}(\lambda)=\text{gamma}(1,\lambda)$. $\mathbb{E}(X)=n\lambda^{-1}$ and $\mathrm{var}(X)=n\lambda^{-2}$.

This also makes sense for real $n>0$ (and $\lambda>0$), if we interpret $(n-1)!$ as $\Gamma(n)$, where $\Gamma(n)=\int_0^\infty x^{n-1}e^{-x}\,dx$.

(d) We say that X is **beta**(a,b) if

$$f_X(x)=\begin{cases} \frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1} & \text{if } 0<x<1,\\ 0 & \text{otherwise.}\end{cases}$$

Here $B(a,b)=\Gamma(a)\Gamma(b)/\Gamma(a+b)$. Then

$$\mathbb{E}(X)=\frac{a}{a+b},\qquad \mathrm{var}(X)=\frac{ab}{(a+b+1)(a+b)^2}.$$

1.10 Laws of large numbers

Suppose $X_1,X_2,\dots$ is a sequence of IID RVs, each having finite mean μ and variance σ^2. Let

$$S_n:=X_1+X_2+\cdots+X_n,\ \text{ so that } \mathbb{E}(S_n)=n\mu, \mathrm{var}(S_n)=n\sigma^2.$$

The **weak law of large numbers** is that for $\epsilon>0$,

$$\mathbb{P}(|S_n/n-\mu|>\epsilon)\to 0,\ \text{ as } n\to\infty\,.$$

The **strong law of large numbers** is that

$$\mathbb{P}(S_n/n\to\mu)=1\,.$$

1.11 The Central Limit Theorem

Suppose $X_1,X_2,\dots$ are as above. Define the **standardized** version S_n^* of S_n as

$$S_n^*=\frac{S_n-n\mu}{\sigma\sqrt{n}},\quad \text{so that } \mathbb{E}(S_n^*)=0, \mathrm{var}(S_n^*)=1.$$

Then for large n, S_n^* is approximately standard Normal: for $a<b$,

$$\lim_{n\to\infty}\mathbb{P}(a\leq S_n^*\leq b)=\Phi(b)-\Phi(a)=\lim_{n\to\infty}\mathbb{P}\left(n\mu+a\sigma\sqrt{n}\leq S_n\leq n\mu+b\sigma\sqrt{n}\right).$$

In particular, for large n,

$$\mathbb{P}(|S_n-n\mu|<1.96\sigma\sqrt{n})\doteqdot 95\%$$

since $\Phi(1.96)=0.0975$ and $\Phi(-1.96)=0.025$.

1.12 Poisson process of rate λ

The Poisson process is used to model a process of arrivals: of people to a supermarket checkout, calls at telephone exchange, etc.

Arrivals happen at times

$$T_1,\ T_1+T_2,\ T_1+T_2+T_3,\ \dots$$

where $T_1,T_2,\dots$ are independent and each exponentially distributed with parameter λ. Numbers of arrivals in disjoint intervals are independent RVs, and the number of arrivals in any interval of length t has the $P(\lambda t)$ distribution. The time

$$S_n=T_1+T_2+\cdots+T_n$$

of the nth arrival has the gamma(n,λ) distribution, and $2\lambda S_n\sim\mathcal{X}_{2n}^2$.